LaTexBlend: Scaling Multi-concept Customized Generation with Latent Textual Blending

Supplementary Material

The supplementary material is organized as follows: In Section A, we provide additional experimental results, including further ablation studies, comparisons of single-concept customized generation, and more comparisons with baseline methods. In Section B, we provide the implementation details of our method and the baselines. In Section C, we discuss the societal impacts of our work.

A. More Experimental Results

A.1. More Ablation Studies

Additional Ablation Study on Base Encoding Flow. We explore whether concept-related information is dispersed across latent textual features when the fine-tuning prompt contains only concept-related tokens. Specifically, the prompt takes the form of "V* <noun>", where "V*" is an identifier token and "<noun>" is a coarse class descriptor of the subject (*e.g.* , cat, dog). To validate the necessity of the base encoding flow in this specific scenario, we conduct additional ablation studies. In the *"w/o base flow*"* scenario, the base flow is removed, and fine-tuning is performed using only the single learnable concept flow $\mathscr{F}_c$. Besides, the fine-tuning prompt adopts the form "V* <noun>" and is padded to a fixed sequence length M. We show sample generations in Fig. 1. As demonstrated, the generated images exhibit significant degradation in concept fidelity in the *"w/o base flow*"* scenario. This result suggests that the obtained $\mathbf{h}_c$ lacks sufficient concept-related information, potentially dispersing into the padding tokens. It further confirms the necessity of the base encoding flow for obtaining an effective concept representation.

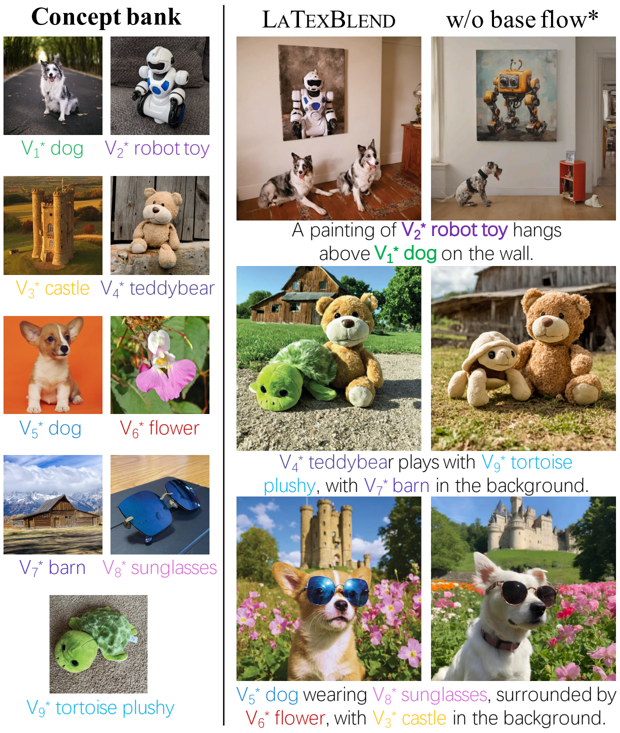

Concept bank
LaTexBlend
w/o base flow*
V1* dog
V2* robot toy
V3* castle
V4* teddybear
V5* dog
V6* flower
V7* barn
V8* sunglasses
V9* tortoise plushy
A painting of V2* robot toy hangs above V1* dog on the wall.
V4* teddybear plays with V9* tortoise plushy, with V7* barn in the background.
V5* dog wearing V8* sunglasses, surrounded by V6* flower, with V3* castle in the background.

Figure 1. **Additional ablation study on the base encoding flow.** The base encoding flow remains necessary when the fine-tuning prompt contains only concept-related tokens. Without it, the obtained $\mathbf{h}_c$ lacks sufficient concept-related information, which may potentially disperse into the padding tokens.

Position Invariance At inference, the blending position of $\mathbf{h}_c$ can vary throughout the prompt. Therefore, we employ a prompt variation strategy for single-concept customization in LaTexBlend, dynamically varying the prompt template to construct textual prompts for the two textual encoding flows. By varying the extraction and insertion positions of $\mathbf{h}_c$, we aim to eliminate its positional dependency. We present sample generations in Fig. 2 to illustrate the position invariance of $\mathbf{h}_c$. Different columns use $\mathbf{h}_c$ obtained from different prompt templates. Specifically, the $\mathbf{h}_c$ used in columns 1, 2, and 3 are extracted from the templates "{}.", "Photo of {}.", and "A fancy photo of {}.", respectively. For each generation case, we present three images generated with different noise initializations. As observed, although the representations $\mathbf{h}_c$ used in different columns are extracted from different templates, they function correctly regardless of their extraction positions and produce similar results with the same noise initialization and prompt. Furthermore, the invariance of $\mathbf{h}_c$ with respect to its extraction position remains robust when its blending position varies. In our experiment, all concept representations $\mathbf{h}_c$ are extracted from the prompt template "Photo of {}.".

Quantitative Ablation on Blending Guidance. We conduct a qualitative ablation study on blending guidance in the paper for an intuitive demonstration. We further perform a quantitative ablation on blending guidance and present the results in Table 1. In the *"w/o blending guidance"* scenario, we merge multiple concepts by performing direct feature replacement in the latent textual space. The straightforward merging strategy performs well when the number of con-

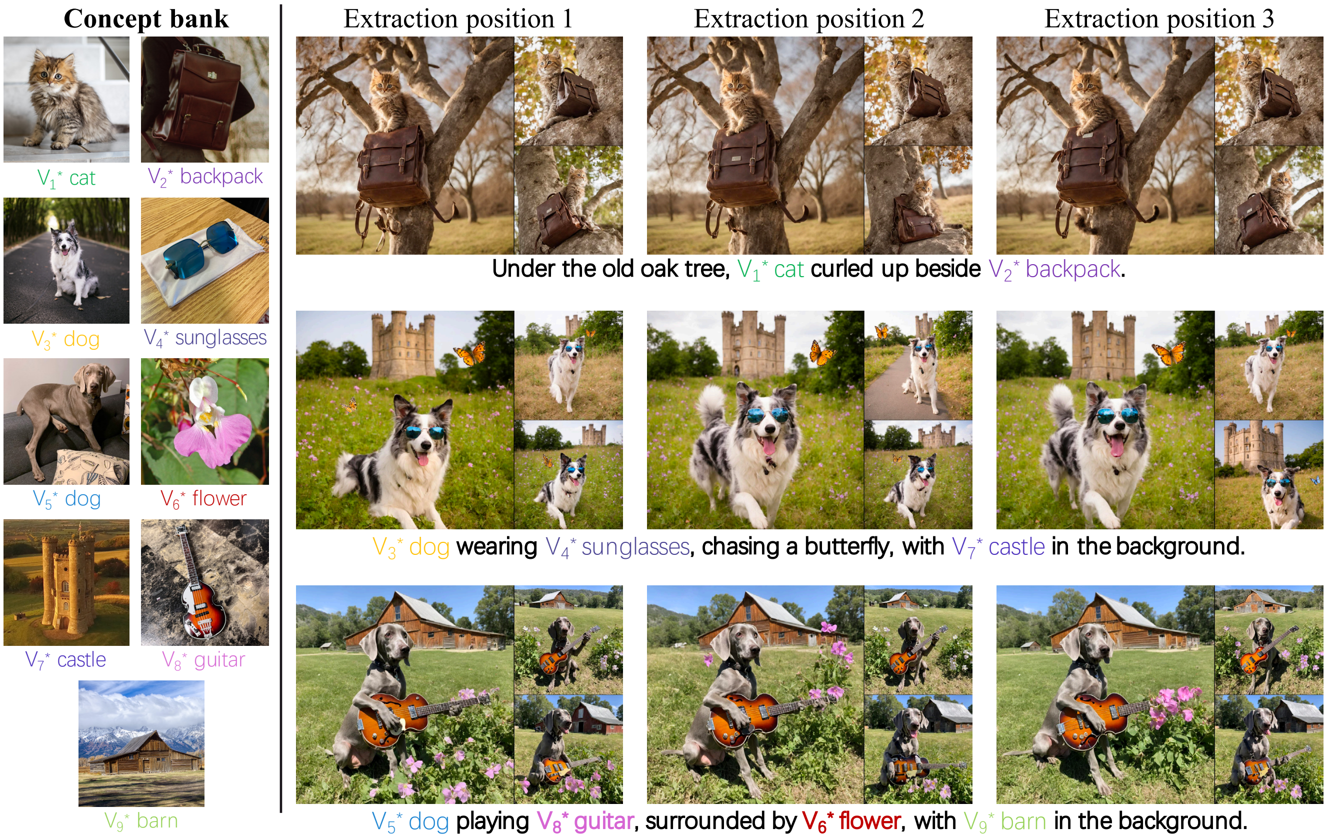

Concept bank
V_1^* cat
V_2^* backpack
V_3^* dog
V_4^* sunglasses
V_5^* dog
V_6^* flower
V_7^* castle
V_8^* guitar
V_9^* barn
Extraction position 1
Extraction position 2
Extraction position 3
Under the old oak tree, V_1^* cat curled up beside V_2^* backpack.
V_3^* dog wearing V_4^* sunglasses, chasing a butterfly, with V_7^* castle in the background.
V_5^* dog playing V_8^* guitar, surrounded by V_6^* flower, with V_9^* barn in the background.

Figure 2. **Position invariance.** Different columns use $\mathbf{h}_c$ extracted from different prompt templates. For each generation case, we present three images generated with different noise initializations. Although the representations $\mathbf{h}_c$ used in different columns are extracted from different templates, they function correctly regardless of their extraction positions and produce similar results with the same noise initialization and prompt.

Variant	$S^{\mathrm{T}}_{\mathrm{CLIP}}$ (↑)	$S^{\mathrm{I}}_{\mathrm{CLIP}}$ (↑)	S_{DINO} (↑)
	`Two concepts`		
w/o blending guidance	0.3286	0.8449	0.6973
LaTexBlend (Ours)	0.3213	0.8495	0.7016
	`Three concepts`		
w/o blending guidance	0.3703	0.806	0.6663
LaTexBlend (Ours)	0.3782	0.8241	0.6913
	`Four concepts`		
w/o blending guidance	0.3831	0.6412	0.5137
LaTexBlend (Ours)	0.4058	0.7419	0.5765

Table 1. **Quantitative Ablation on Blending Guidance.** Blending guidance offers a subtle improvement with fewer concepts, while enhancing multi-concept generation as the number of concepts increases.

cepts is small, with blending guidance offering only subtle improvement. As the number of concepts increases, blending guidance greatly enhances multi-concept inference.

A.2. Language Expression Ability

Regarding the generation quality of multi-concept generation, apart from high layout coherence and concept fidelity, another key advantage of LaTexBlend is its strong language expression ability. Our method alleviates the decline in language expression ability observed in previous customized generation methods. We present sample results in Fig. 3. LaTexBlend effectively preserves the editability of the pre-trained model on customized concepts, allowing for flexible modifications to the material, actions, and colors of the customized concepts.

A.3. Generations of Similar Concepts

In some multi-concept generation cases, the coarse class descriptor of the subject may be the same. The merit of the latent textual space and blending guidance ensure that multiple similar customized concepts sharing the same common class descriptors can be generated together with high image quality. We present sample generations in Fig. 4. As we can see, LaTexBlend can generate customized concepts

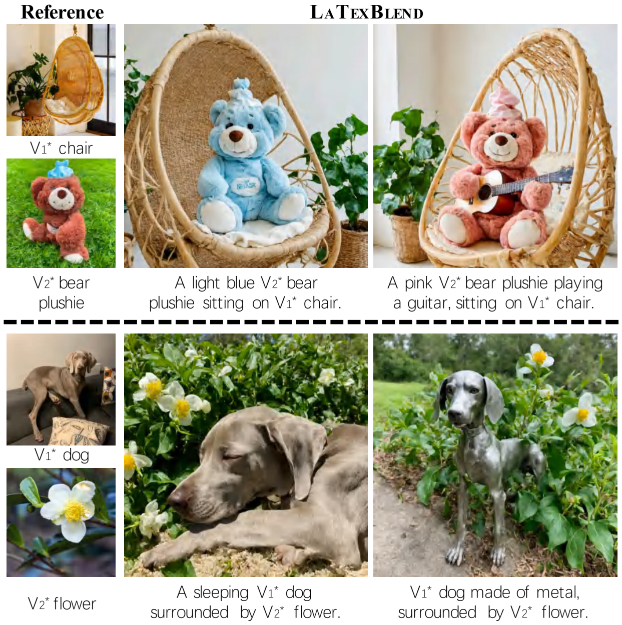

Reference
LaTexBlend
V1* chair
V2* bear plushie
A light blue V2* bear plushie sitting on V1* chair.
A pink V2* bear plushie playing a guitar, sitting on V1* chair.
V1* dog
V2* flower
A sleeping V1* dog surrounded by V2* flower.
V1* dog made of metal, surrounded by V2* flower.

Figure 3. **Language expression ability on customized concepts.** LATEXBLEND effectively preserves the editability of the pre-trained model on customized concepts, allowing for flexible modifications to the material, actions, and colors of the customized concepts.

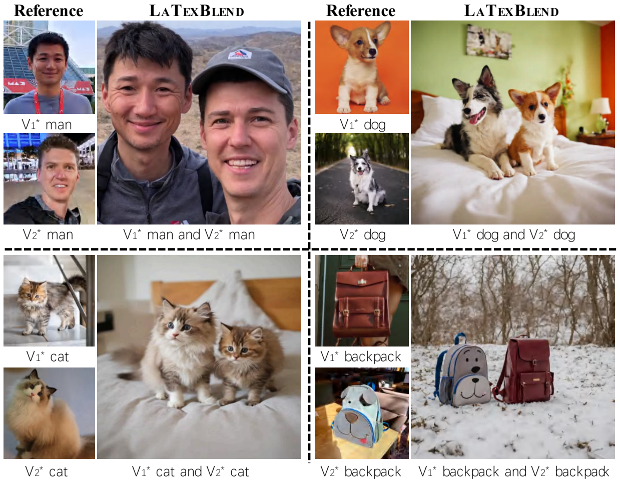


Figure 4. **Generations of Similar Concepts.** LATEXBLEND can generate customized concepts that share the same common words within the same image without identity confusion.

that share the same common words, such as "dog" or "man", within the same image with high concept fidelity and strong prompt adherence.

A.4. Single-concept Customized Generation

The proposed LATEXBLEND can also be applied to single-concept customized generation, where a single concept representation is blended with the output of the base encoding flow in the latent textual space:

$$\mathcal{F}(\mathbf{h}_{\mathcal{C}}) = \texttt{Blend}(\mathbf{h}_b; \mathbf{h}_{c_1})\,, \tag{1}$$

where $\mathbf{h}_b$ denotes the output of the base encoding flow, while $\mathbf{h}_{c_1}$ is the representation of the target concept. We compare LATEXBLEND with several representative single-concept customized generation methods, including DreamBooth [14], Custom Diffusion [11], and LoRA [6]. Sample generations are presented in Fig. 5. As observed, DreamBooth fine-tunes the entire U-Net, which often leads to overfitting and a loss of editability, making it difficult to accurately render the target subject within the query context. Custom Diffusion falls short in concept fidelity. In contrast, LATEXBLEND generates the customized subject with high concept fidelity while faithfully adhering to the query prompt. This further validates the effectiveness of LATEXBLEND in representing concepts and mitigating denoising deviation in customized generation.

A.5. Comparison with More Baselines

Due to space limitations, we omit visual comparison with some earlier multi-concept customized generation methods in the paper. Therefore, we supplement the qualitative comparisons of LATEXBLEND with these methods, including Cones 2 [12] and Custom Diffusion [11]. For each method, we randomly generate 10 images per case and select the best 3 for visual comparison. We present sample generations in Fig. 6. As observed, Cones 2 relies heavily on additional layout conditioning. In the absence of predefined layout conditions, Cones 2 suffers from degradation in both subject fidelity and image structure. Images generated with explicit layout conditions lack diversity in their layouts and may struggle to capture complex semantics, such as inter-subject interactions and actions. Custom Diffusion falls short in maintaining subject fidelity and ensuring coherence in image structure.

Guidance [1, 5] is a commonly used technique in conditional image generation, which adjusts the sampling process toward specific targets by modifying the update rule of noisy latents. We propose blending guidance in LATEXBLEND, which leverages mutual information from different concepts within a single denoising branch to rectify attention. MC^2 [8] is a guidance-based multi-concept customized generation method. Unlike LATEXBLEND, MC^2 derives guidance from attention relations across multiple denoising branches. Therefore, its inference-time computation is high and scales proportionally with the number of concepts.

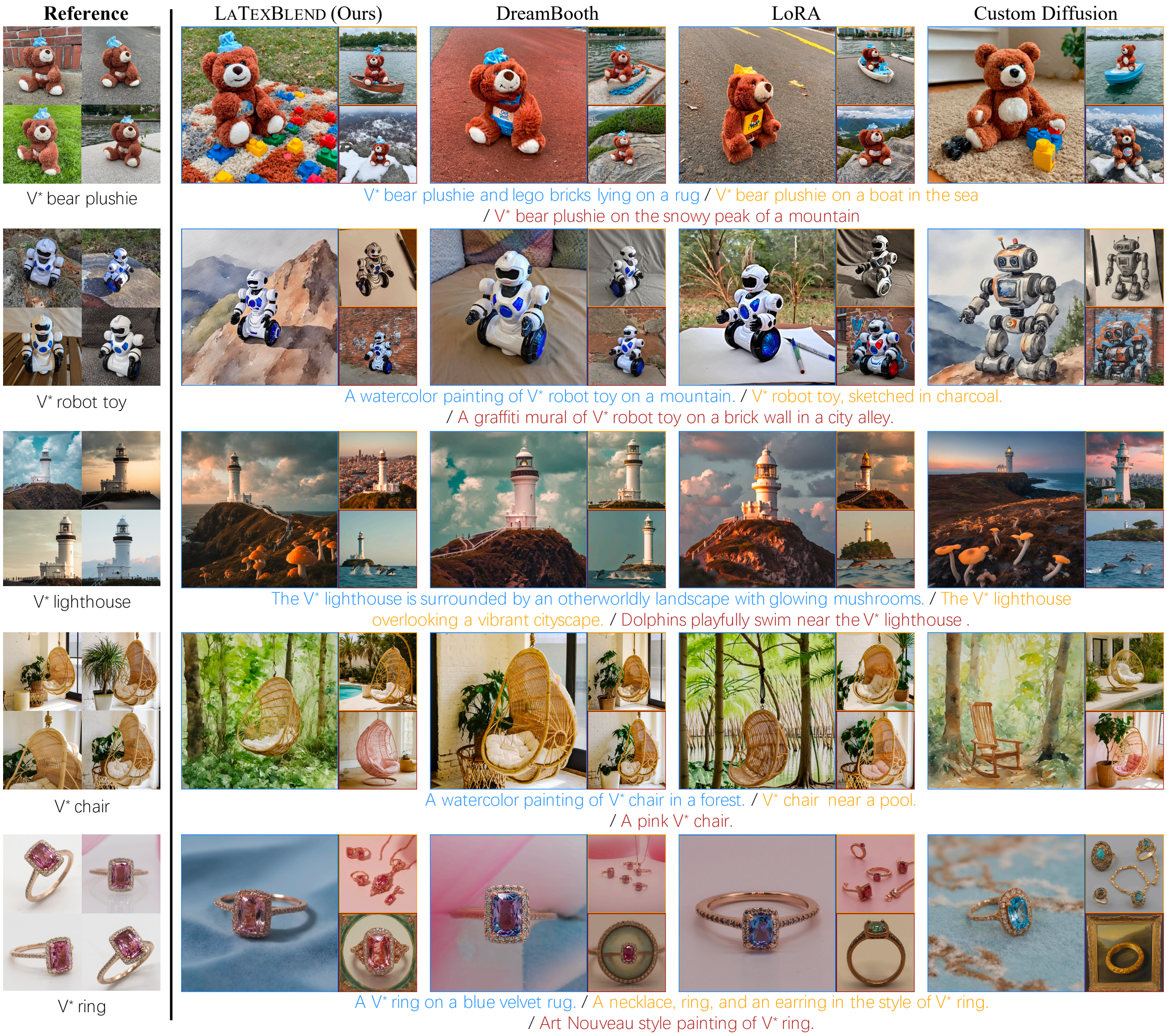

Reference
LATEXBLEND (Ours)
DreamBooth
LoRA
Custom Diffusion
V* bear plushie
V* bear plushie and lego bricks lying on a rug / V* bear plushie on a boat in the sea
/ V* bear plushie on the snowy peak of a mountain
V* robot toy
A watercolor painting of V* robot toy on a mountain. / V* robot toy, sketched in charcoal.
/ A graffiti mural of V* robot toy on a brick wall in a city alley.
V* lighthouse
The V* lighthouse is surrounded by an otherworldly landscape with glowing mushrooms. / The V* lighthouse
overlooking a vibrant cityscape. / Dolphins playfully swim near the V* lighthouse .
V* chair
A watercolor painting of V* chair in a forest. / V* chair near a pool.
/ A pink V* chair.
V* ring
A V* ring on a blue velvet rug. / A necklace, ring, and an earring in the style of V* ring.
/ Art Nouveau style painting of V* ring.

Figure 5. **Visual comparison of single-concept customized generation.** We set the rank of LoRA to 4 by default. DreamBooth often overfits, resulting in a loss of editability, while Custom Diffusion struggles to maintain concept fidelity. LATEXBLEND generates the customized subject with high concept fidelity while faithfully adhering to the query prompt.

A.6. Comparison with Learning-based Method

Apart from optimization-based methods, another line of research on customized generation is learning-based method [2, 16–18], which aims to train unified models capable of personalizing diverse subject inputs. We compare our approach with several existing learning-based multi-concept customized generation methods, including FreeCustom [2], SSR-Encoder [18], and MS Diffusion [16]. We utilize the official implementation for FreeCustom[1], SSR-Encoder[2], and MS Diffusion[3]. For each competing method, we randomly generate 10 images per case and select the best 3 for visual comparison. The results of the qualitative comparison are shown in Fig. 7. As we can see, these learning-based methods face challenges in faithfully generating target subjects and preserving their key identifying features, resulting in low subject fidelity - especially for complex customized subjects.

[1] https://github.com/aim-uofa/FreeCustom

[2] https://github.com/Xiaojiu-z/SSR_Encoder

[3] https://github.com/MS-Diffusion/MS-Diffusion

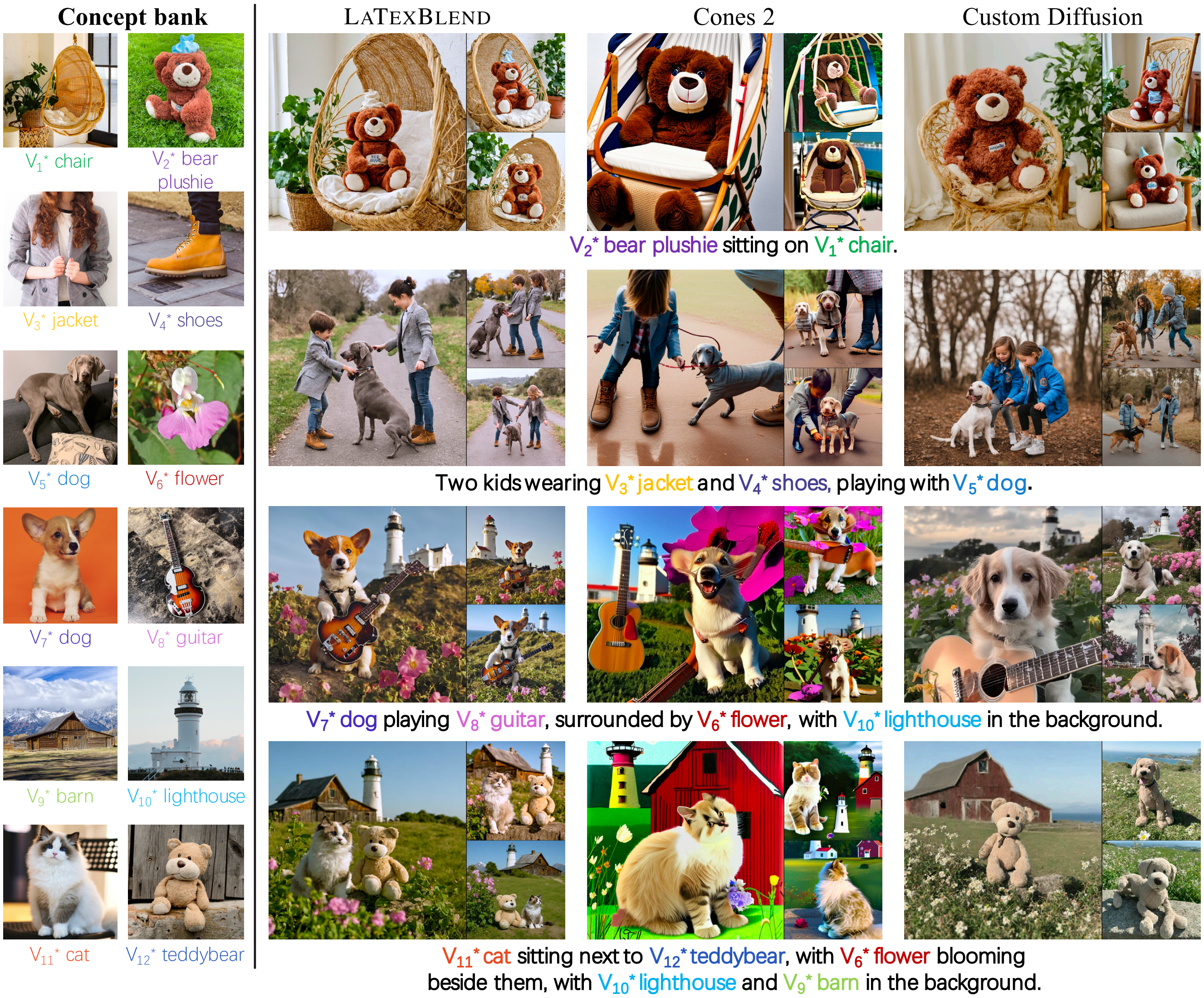

Concept bank
LaTeXBlend
Cones 2
Custom Diffusion
V_1^* chair
V_2^* bear plushie
V_3^* jacket
V_4^* shoes
V_5^* dog
V_6^* flower
V_7^* dog
V_8^* guitar
V_9^* barn
V_{10}^* lighthouse
V_{11}^* cat
V_{12}^* teddybear
V_2^* bear plushie sitting on V_1^* chair.
Two kids wearing V_3^* jacket and V_4^* shoes, playing with V_5^* dog.
V_7^* dog playing V_8^* guitar, surrounded by V_6^* flower, with V_{10}^* lighthouse in the background.
V_{11}^* cat sitting next to V_{12}^* teddybear, with V_6^* flower blooming
beside them, with V_{10}^* lighthouse and V_9^* barn in the background.

Figure 6. **Comparison with more multi-concept customized generation Methods.** We perform qualitative comparisons of LaTeXBlend with two earlier multi-concept customized generation methods, including Cones 2 [12] and Custom Diffusion [11]. LaTeXBlend demonstrates advantages over the baselines in both subject fidelity and image structure coherence.

A.7. Supplementary Experimental Results

More Results of Visual Comparison. We provide more qualitative comparisons of multi-concept generation between LaTeXBlend and baseline methods, including Custom Diffusion [11], Mix-of-Show [4], OMG [10], and MuDI [7], as shown in Fig. 8. For each method, we randomly generate 10 images per case and select the best 3 for visual comparison. MuDI, Mix-of-Show, and Custom Diffusion exhibit issue of image structure degradation, often producing single-object-centric subjects or omitting target subjects. OMG's performance heavily relies on the accuracy of the segmentation model [9], occasionally failing to integrate customized subjects, which results in gray-shaded areas. Besides, OMG struggles to maintain overall consistency and realism in image style.

Detailed Data of Quantitative Comparison. The detailed data of Fig. 7 in the paper is provided in Table 2. As shown, the proposed LaTeXBlend significantly outperforms all baseline methods in concept alignment, particularly in terms of the DINO score. Compared with CLIP, DINO can better capture the unique features of each subject, thereby better reflecting fine subject similarity rather than coarse class similarity [14]. The superiority in DINO score highlights LaTeXBlend's ability to effectively pre-

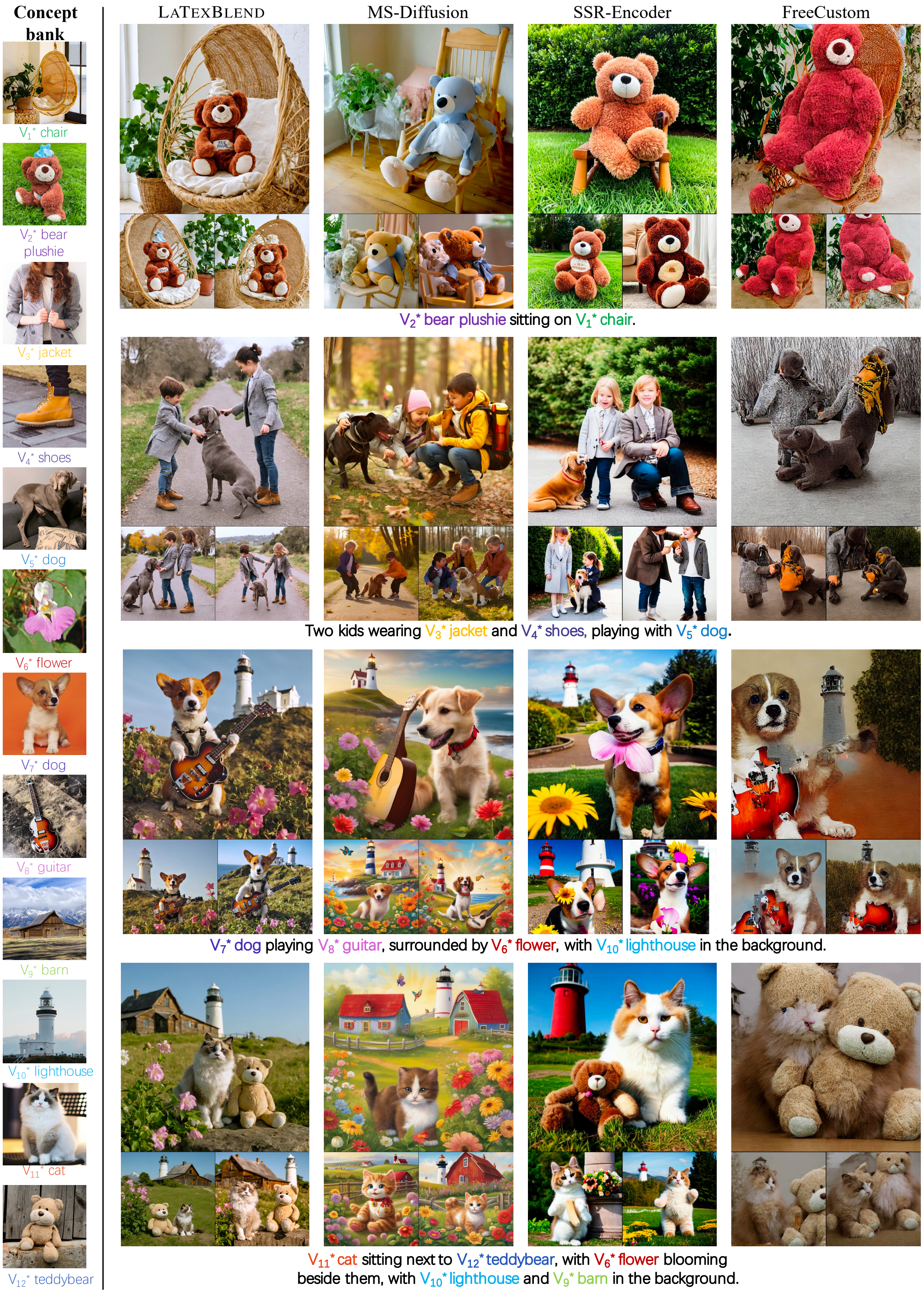

Concept bank
LaTexBlend
MS-Diffusion
SSR-Encoder
FreeCustom
V_1^* chair
V_2^* bear plushie
V_3^* jacket
V_4^* shoes
V_5^* dog
V_6^* flower
V_7^* dog
V_8^* guitar
V_9^* barn
V_{10}^* lighthouse
V_{11}^* cat
V_{12}^* teddybear
V_2^* bear plushie sitting on V_1^* chair.
Two kids wearing V_3^* jacket and V_4^* shoes, playing with V_5^* dog.
V_7^* dog playing V_8^* guitar, surrounded by V_6^* flower, with V_{10}^* lighthouse in the background.
V_{11}^* cat sitting next to V_{12}^* teddybear, with V_6^* flower blooming beside them, with V_{10}^* lighthouse and V_9^* barn in the background.

Figure 7. **Comparison with learning-based multi-concept customized generation methods.** We compare our approach with several existing multi-concept learning-based methods, including FreeCustom [2], SSR-Encoder [18], and MS Diffusion [16]. The primary issue with learning-based methods is their struggle to generate customized subjects with high subject fidelity, particularly for complex subjects.

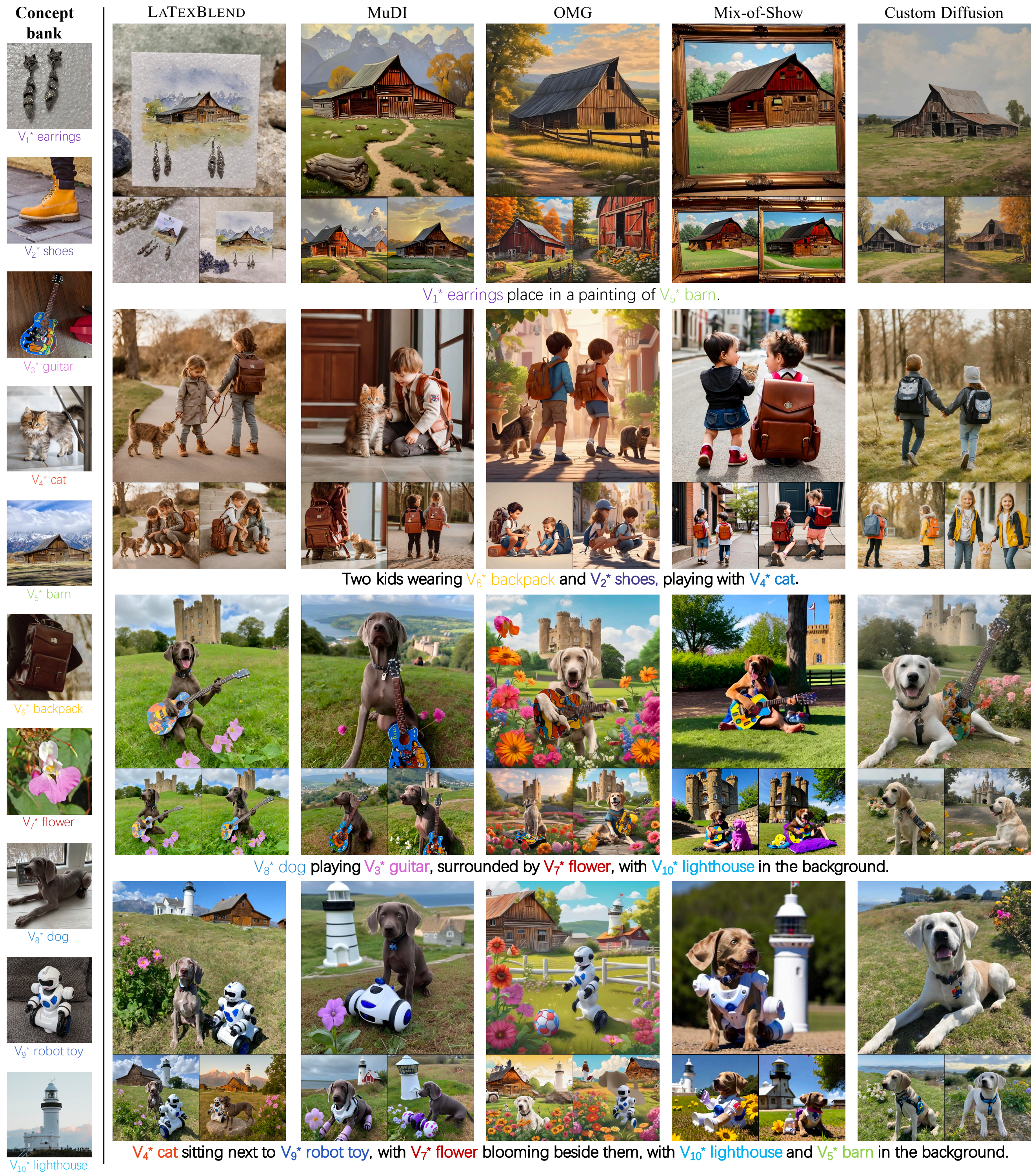

Concept bank
LaTexBlend
MuDI
OMG
Mix-of-Show
Custom Diffusion
V_1^* earrings
V_2^* shoes
V_3^* guitar
V_4^* cat
V_5^* barn
V_6^* backpack
V_7^* flower
V_8^* dog
V_9^* robot toy
V_{10}^* lighthouse
V_1^* earrings place in a painting of V_5^* barn.
Two kids wearing V_6^* backpack and V_2^* shoes, playing with V_4^* cat.
V_8^* dog playing V_3^* guitar, surrounded by V_7^* flower, with V_{10}^* lighthouse in the background.
V_4^* cat sitting next to V_9^* robot toy, with V_7^* flower blooming beside them, with V_{10}^* lighthouse and V_5^* barn in the background.

Figure 8. **More visual comparisons with baselines.** We provide more qualitative comparisons of multi-concept generation between LaTexBlend and baseline methods, including Custom Diffusion [11], Mix-of-Show [4], OMG [10], and MuDI [7].

serve the key identifying features of target subjects.

	Methods	One concept	Two concepts	Three concepts	Four concepts	Five concepts	**Average**	Five concepts w layout
Concept-alignment CLIP score	**Cones 2**	0.7124	0.7079	0.737	0.7219	0.6988	0.7155	0.7614
	Mix-of-Show	0.7222	0.6785	0.6846	0.6766	0.7343	0.6992	0.7067
	OMG	0.7786	0.7405	0.6834	0.7067	0.6974	0.7213	-
	MuDI	0.7415	0.7131	0.7232	0.7379	0.7641	0.7359	0.7211
	LATEXBLEND (Ours)	0.7766	0.7322	0.7688	0.7553	0.795	**0.7656**	**0.7829**
Concept-alignment DINO score	**Cones 2**	0.3494	0.3903	0.4151	0.4286	0.363	0.3893	0.4278
	Mix-of-Show	0.4838	0.4817	0.4055	0.395	0.3915	0.4315	0.3886
	OMG	0.4914	0.5213	0.4994	0.4405	0.4874	0.488	-
	MuDI	0.5345	0.4902	0.4995	0.5002	0.488	0.5025	0.3826
	LATEXBLEND (Ours)	0.5846	0.5892	0.5729	0.4922	0.5196	**0.5517**	**0.5214**
Text-alignment	**Cones 2**	0.3857	0.3201	0.3205	0.2959	0.3899	0.3424	0.3193
	Mix-of-Show	0.334	0.3101	0.2817	0.3224	0.3893	0.3275	**0.353**
	OMG	0.3616	0.3433	0.3215	0.4299	0.3816	**0.3675**	-
	MuDI	0.35	0.2817	0.3035	0.3805	0.3687	0.3368	0.3109
	LATEXBLEND (Ours)	0.3745	0.327	0.3025	0.4242	0.4044	0.3665	0.348

Table 2. **Detailed data of quantitative evaluation on multi-concept generation**. We highlight the best result in bold and underline the second best for different settings. LATEXBLEND outperforms all baseline methods in concept alignment and demonstrates competitive performance in prompt fidelity. The clear advantages of LATEXBLEND in the DINO score showcase its ability to effectively preserve the key identifying features of target subjects.

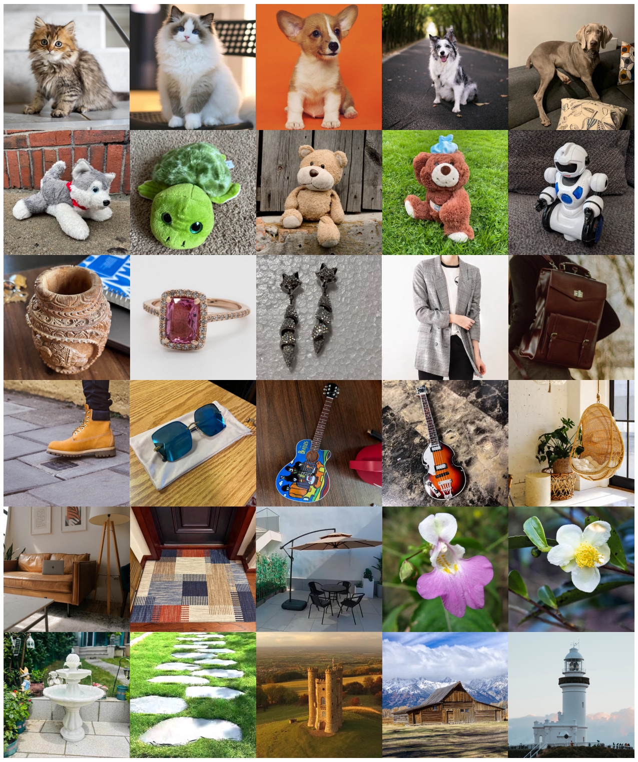


Figure 9. **One reference image of each subject.**

B. Implementation Details

B.1. Subjects and Prompts

We conduct experiments on 30 different subjects, most of which are sourced from previous studies [3, 11, 14]. Additionally, we collect images of some new subjects missing in previous studies. These subjects cover various categories, such as buildings, pets, and objects, each represented by several reference images. We show one reference image of each subject in Fig. 9. The subject combinations and corresponding query prompts are initially generated by ChatGPT and subsequently reviewed and curated manually. Specifically, we create a `concept list` containing the 30 subjects mentioned above, then generate n-concept subject combinations and their query prompts using ChatGPT with the following several steps (where n is replaced by the specific number): 1. Select n subjects from the `concept list`. 2. Use the selected n nouns to construct sentences based on the following requirements: 1) Apart from the n selected nouns, avoid using other nouns from the concept list in the sentence. 2) Ensure the sentences are logical. 3) Maintain diversity in the sentence

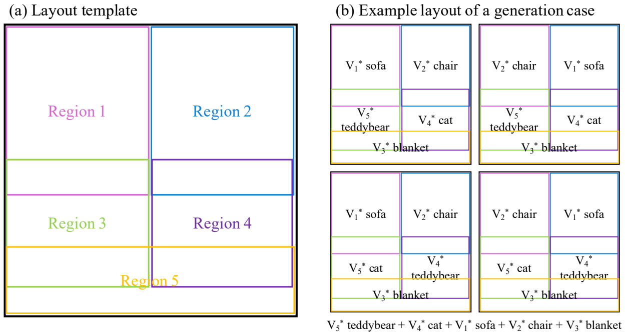


Figure 10. **Layout template for layout-conditioned multi-concept generation. (a).** The layout template. **(b).** For each generation case, we sequentially swap the specified subjects in regions 1 and 2, and those in regions 3 and 4, resulting in a total of 4 different layout instances.

structures. Some prompts are also inspired by other previous works [11, 14].

B.2. Prompt Template Pool

For the customization of each single concept in LaTexBlend, we create a template pool containing 7 different prompt templates. The complete list of templates is:

1. "{}."
2. "A {}."
3. "Photo of {}."
4. "A photo of {}."
5. "A photo of a {}."
6. "a fancy photo of a {}."
7. "A fancy, detailed photo of {}."

During fine-tuning, we randomly draw different templates from the prompt template pool to construct prompts for the two textual encoding flows. At inference, the compact representation $\mathbf{h}_c$ of each concept is obtained from the template "Photo of {}.".

B.3. Layout Conditioning

We conduct experiments of multi-concept customized generation with additional layout conditioning in Section 4.2 of the paper. The layout template we use is shown in Fig. 10 (a). There are 5 specified generation regions in the layout template for 5-concept generation. For each generation case, we alternately swap the specified subjects in regions 1 and 2, and those in regions 3 and 4, resulting in a total of 4 different layout instances. We also provide an example of all 4 layout instances for a generation case in Fig. 10 (b). For each generation case, we randomly generate 5 images per layout instance using different methods and select the best 3 from the resulting 20 images for visual comparison.

B.4. Additional Details on User Study

We conduct a user study with 25 participants, using 20 sets of generation cases with the number of concept ranging from 2 to 5. For each generation case, we randomly generate 5 images per query prompt using each competing method to create an image candidate pool. Before evaluation, participants are thoroughly briefed on the scoring rules and provided with scoring examples. We provide a screenshot of the instructions given to participants in Fig. 11. The images from Cones 2 in the user study are generated without additional layout conditioning for a fair comparison. In each evaluation case, participants are given a textual prompt, reference images of the customized subject, and corresponding generations from different methods. Generated images are randomly selected from the image candidate pool and presented side-by-side in a random order to participants. Participants are given unlimited time to score each generation on a scale from 0 to 5 (with 0 being the worst and 5 the best) based on three criteria: 1) whether the image contains all target subjects and aligns with their visual appearance in the reference images, 2) whether the image content adheres to the scenes described by the textual prompt, and 3) the overall quality in terms of authenticity and coherence. A screenshot of an evaluation case is provided in Fig. 12.

B.5. Implementation Details

In our experiment, we use Stable Diffusion XL (SDXL) [13] as the pre-trained text-to-image diffusion model. Images are generated using 100 DDIM sampling steps with a classifier-free guidance scale of 6 for all compared methods. All model fine-tuning is conducted on NVIDIA GeForce RTX 4090 GPUs, and inference is conducted on a 40GB NVIDIA A100 GPU.

Custom Diffusion. We employ the official implementation[4] for Custom Diffusion [11]. The model is fine-tuned using the default hyperparameters and settings provided in the code. Custom Diffusion requires joint training for multi-concept generation. The recommended fine-tuning steps are 500 for a single concept and 1000 for two concepts; accordingly, we increase the training steps by 500 for each additional concept.

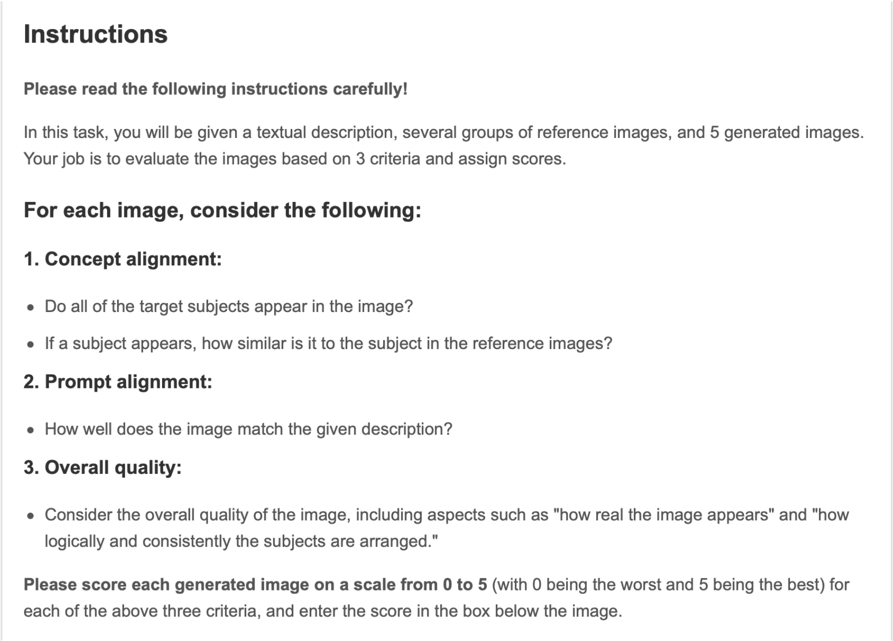
Mix-of-Show. We use the official implementation[5] of Mix-of-Show [4]. Following the authors' guidelines and examples, we make extra annotations for the reference images, including subject masks and detailed image captions. For single-concept fine-tuning, each concept is associated with two identifier tokens, represented in the form "[V1] [V2] <noun1>". We fine-tune the single-concept model and fuse multiple models using the default parameters provided in the official implementation. The fusion operation is performed for each distinct subject combination. For multi-concept generation with layout conditions, we manually create sketch-based conditions in the form of precise object contours, following the format of the official examples.

Cones 2. The official implementation[6] of Cones 2 [12] uses Stable Diffusion V2.1 as its pre-trained diffusion model. For a fair comparison, we upgrade the backbone in the official implementation to SDXL. We provide sample results generated by the original implementation in Fig. 13 for reference. In the comparison of multi-concept generation without additional layout conditioning, explicit layout guidance is omitted.

[4] https://github.com/adobe-research/custom-diffusion

[5] https://github.com/TencentARC/Mix-of-Show

[6] https://github.com/ali-vilab/Cones-V2



Instructions

Please read the following instructions carefully!

In this task, you will be given a textual description, several groups of reference images, and 5 generated images. Your job is to evaluate the images based on 3 criteria and assign scores.

For each image, consider the following:

1. Concept alignment:

- Do all of the target subjects appear in the image?
- If a subject appears, how similar is it to the subject in the reference images?

2. Prompt alignment:

- How well does the image match the given description?

3. Overall quality:

- Consider the overall quality of the image, including aspects such as "how real the image appears" and "how logically and consistently the subjects are arranged."

Please score each generated image on a scale from 0 to 5 (with 0 being the worst and 5 being the best) for each of the above three criteria, and enter the score in the box below the image.

Figure 11. **The instructions that were given to the participants.**

OMG. We employ the official implementation[7] of OMG [10]. The segmentation model `SAM+Grounding-DINO` is used to generate concept masks. Following the authors' recommendation, we first train single-concept LoRA models using the code provided in the repository[8]. To enhance the LoRA model's ability to capture key identifying features of concepts, we increase the rank dimensionality from 4 to 20. During fine-tuning, the textual prompt for reference images is formatted as `"[V] <noun>"`.

MuDI. We utilize the official implementation[9] of MuDI [7]. Following the authors' guidelines and examples, we make extra annotations for all reference images, including subject masks and detailed image captions. The model is fine-tuned using the default parameters from the official implementation for a total of 2000 steps. Samples are generated using model checkpoints at steps 400, 600, 800, 1000, and 2000, with the best outputs selected for comparison. For generation without layout conditions, we follow the official protocol, employing latent initialization with random position. For generation with explicit layout conditions, we manually specify the ordering and positioning of object initializations, incorporating masks as conditions into the initialization of latent variables.

LaTexBlend (Ours). Our code is implemented based on the diffusers library [15]. Our method does not require additional annotations of reference images for fine-tuning. The prompt template we used for fine-tuning is described in Section B.2. We utilize real images as the regularization dataset, with a prior loss weight of 1.0. The images are augmented using `RandomHorizontalFlip` with a flip probability of 0.5. The prompts for regularization images follow the format `"A <noun>"`. The model is fine-tuned with a batch size of 1 over 500 steps for single-concept customization. The base learning rate for both model parameters and concept embeddings is set to 10^{-5}. After fine-tuning, concept representations are extracted from the prompt template `"Photo of {}."` for all concepts, as detailed in Section A.1.

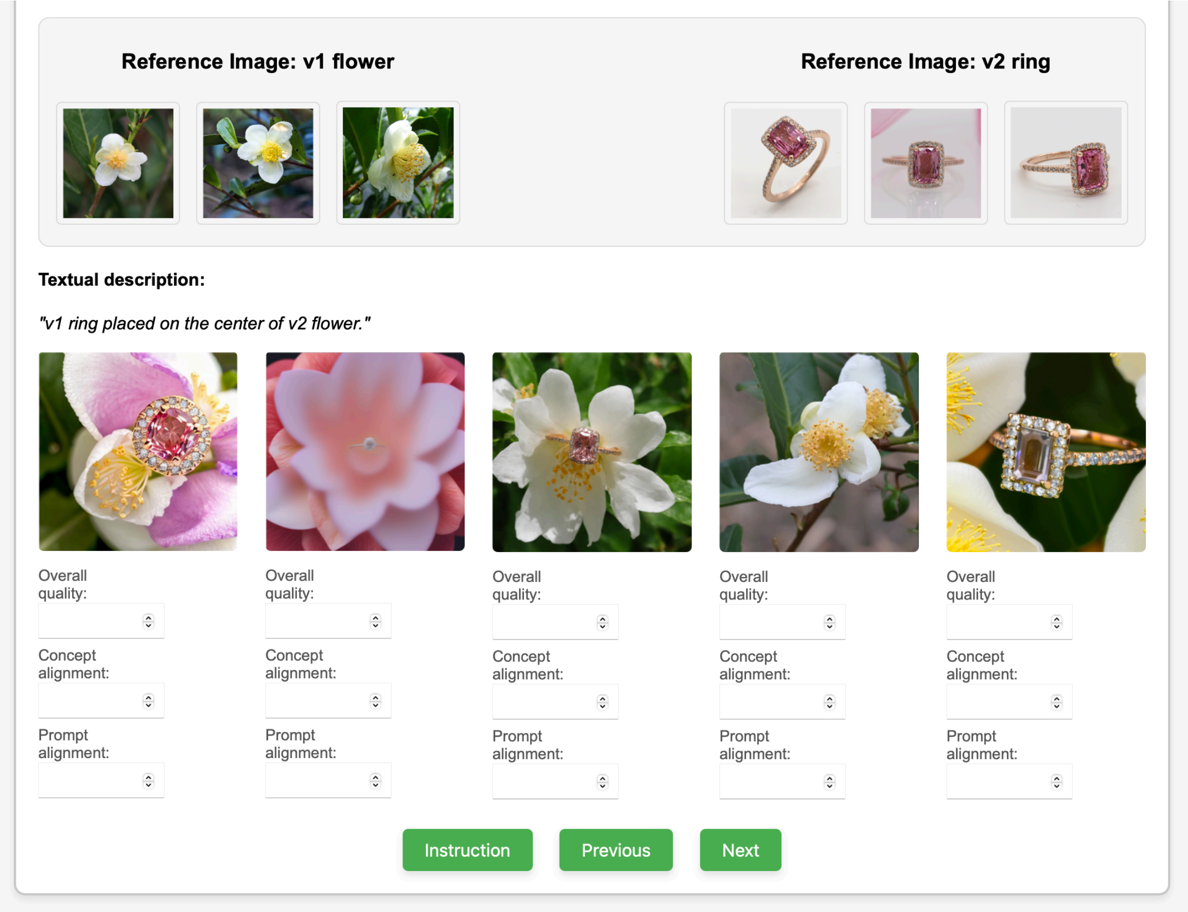
C. Societal Impact

Our method can seamlessly integrate multiple customized subjects into a single image with high quality while maintaining computational efficiency. This advancement democratizes access to high-quality text-to-image generation technologies, offering greater flexibility and personalization for customized content creation and enabling broader applications across creative industries. However, the potential misuse of such technologies, including the generation of misleading or harmful content, raises ethical concerns. Recent research on safeguards, such as reliable detection methods for fake generated data, provides a promising approach to mitigating these potential negative impacts.

[7] https://github.com/kongzhecn/OMG

[8] https://github.com/kohya-ss/sd-scripts

[9] https://github.com/agwmon/MuDI



Reference Image: v1 flower

Reference Image: v2 ring

Textual description:

"v1 ring placed on the center of v2 flower."

Overall quality:	Overall quality:	Overall quality:	Overall quality:	Overall quality:
Concept alignment:	Concept alignment:	Concept alignment:	Concept alignment:	Concept alignment:
Prompt alignment:	Prompt alignment:	Prompt alignment:	Prompt alignment:	Prompt alignment:

Figure 12. **A screenshot of an evaluation case from the user study.**

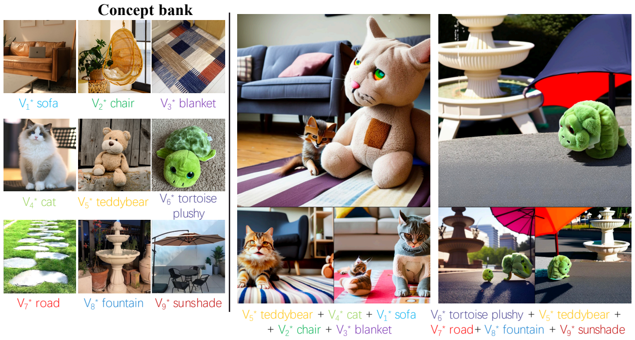

Concept bank
V_1^* sofa
V_2^* chair
V_3^* blanket
V_4^* cat
V_5^* teddybear
V_6^* tortoise plushy
V_7^* road
V_8^* fountain
V_9^* sunshade
V_5^* teddybear + V_4^* cat + V_1^* sofa + V_2^* chair + V_3^* blanket
V_6^* tortoise plushy + V_5^* teddybear + V_7^* road+ V_8^* fountain + V_9^* sunshade

Figure 13. **Sample generations of Cones 2 with Stable Diffusion V2.1.** The generated images exhibit issues such as concept omission and low concept fidelity.

References

[1] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3

[2] Ganggui Ding, Canyu Zhao, Wen Wang, Zhen Yang, Zide Liu, Hao Chen, and Chunhua Shen. Freecustom: Tuning-free customized image generation for multi-concept composition. In *CVPR*, pages 9089–9098, 2024. 4, 6

[3] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 8

[4] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 5, 7, 9

[5] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[6] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 3

[7] Sangwon Jang, Jaehyeong Jo, Kimin Lee, and Sung Ju Hwang. Identity decoupling for multi-subject personalization of text-to-image models. In *NeurIPS*, 2024. Accepted. 5, 7, 10

[8] Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, Wenbo Li, Renjing Pei, Fan Li, and Wangmeng Zuo. Mc^2: Multi-concept guidance for customized multi-concept generation. *arXiv preprint arXiv:2404.05268*, 2024. 3

[9] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 5

[10] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*, 2024. 5, 7, 10

[11] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, pages 1931–1941, 2023. 3, 5, 7, 8, 9

[12] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *NeurIPS*, pages 57500–57519, 2023. 3, 5, 9

[13] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 9

[14] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 3, 5, 8, 9

[15] Pierre von Platen, Suraj Patil, Andrey Lozhkov, Pablo Cuenca, Nicolas Lambert, Kashif Rasul, Munkhdalai Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. `https://github.com/huggingface/diffusers`, 2022. 10

[16] X Wang, Siming Fu, Qihan Huang, Wanggui He, and Hao Jiang. Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. *arXiv preprint arXiv:2406.07209*, 2024. 4, 6

[17] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, pages 15943–15953, 2023.

[18] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al. Ssr-encoder: Encoding selective subject representation for subject-driven generation. In *CVPR*, pages 8069–8078, 2024. 4, 6